Andmete puhastamine, töötlemine

https://towardsdatascience.com/tutorial-data-wrangling-and-mapping-in-r-ec828acc8073 https://programminghistorian.org/en/lessons/data_wrangling_and_management_in_R https://rpubs.com/bradleyboehmke/data_wrangling

https://rpubs.com/bradleyboehmke/data_processing https://suzan.rbind.io/2018/01/dplyr-tutorial-1/

Andmed

Laeme andmed

```
us_state_populations<-read.csv("data/introductory_state_example.csv")
head(us_state_populations)
```

```
##   year         state population
## 1 1790   Connecticut     237655
## 2 1790      Delaware      59096
## 3 1790       Georgia      82548
## 4 1790      Maryland     319728
## 5 1790 Massachusetts     475199
## 6 1790 New Hampshire     141899
```

Vaatame veel, mis andmetes on.

```
str(us_state_populations)
```

```
## 'data.frame':    981 obs. of  3 variables:
##  $ year      : int  1790 1790 1790 1790 1790 1790 1790 1790 1790 1790 ...
##  $ state     : Factor w/ 83 levels "Alabama","Alaska",..: 11 13 17 34 35 48 49 52 55 64 ...
##  $ population: int  237655 59096 82548 319728 475199 141899 184139 340241 395005 433611 ...
```

```
summary(us_state_populations)
```

```
##       year                 state       population
##  Min.   :1790   Connecticut  : 23   Min.   :    4762
##  1st Qu.:1870   Delaware     : 23   1st Qu.:  412198
##  Median :1920   Georgia      : 23   Median : 1131597
##  Mean   :1916   Maryland     : 23   Mean   : 2440995
##  3rd Qu.:1970   Massachusetts: 23   3rd Qu.: 2915841
##  Max.   :2010   New Hampshire: 23   Max.   :37253956
##                 (Other)      :843
```

```
dim(us_state_populations)
```

```
## [1] 981   3
```

tidyverse

Nüüd jõuab kätte osa, mis on üks oluline R-i populaarsuse faktor. **Tidyverse** on andmeteaduse pakettide kogum. Üheks autoriks on Hadley Wickham, R-i arendajate seas legendaarne kuju.

Tidyverse paketid saab alla laadida ühe korraga:

```
install.packages("tidyverse")
```

Figure 1: Hadley Wickham

Järgnevalt mõned näited tidyverse'i kogumist.

Filtreerimine

Ütleme, et tahame filterdada välja read, mis käivad California ja New York'i kohta.

```
library(tidyverse)
#tulemus omistatakse muutujale df_california_ny, aluseks võtame andmed
#muutujast us_state_populations
df_california_ny<-us_state_populations %>%
  #filtreerimme välja osariigid, millen nimed on alltoodud vekto  ris c(...)
  filter(state %in% c("California", "New York"))
dim(df_california_ny)
```

```
## [1] 40  3
```

```
head(df_california_ny)
```

```
##   year    state population
## 1 1790 New York     340241
## 2 1800 New York     589051
## 3 1810 New York     959049
## 4 1820 New York    1372812
## 5 1830 New York    1918608
## 6 1840 New York    2428921
```

Eelnevalt me kasutasime pipe oprtaatorit (%>%). See on väga mugav vahend, mitme operatsiooni tegemiseks. Näiteks me tahame leida keskmist elanike arvu Californias ja New Yorkis kogu vaadeldava perioodi kohta (st iga aasta on kaaluga 1).

Ilma pipe operaatorita näeb see välja nii:

```
us_state_populations %>%
  #filtreerimme välja osariigid, millen nimed on alltoodud vektoris c(...)
  filter(state %in% c("California", "New York"))%>%
  #grupeerime osariigi järgi
  group_by(state)%>%
  #arvutame iga grupi kohta keskmise ja st. hälve
  summarise(mean=mean(population), sandardhälve=sd(population))
```

```
## # A tibble: 2 x 3
```

```
##   state           mean sandardhälve
##   <fct>          <dbl>        <dbl>
## 1 California 11399403.    12824124.
## 2 New York    8984252.     6900774.
```

Samuti saame pipe kasutada, et saada ülevaade mingist osast andmetest:

```
#populatsioon alates aastast 1900
us_state_populations %>%
  select(year, population) %>%
  filter(year>=1900)%>%
  glimpse()
```

```
## Observations: 614
## Variables: 2
## $ year       <int> 1900, 1900, 1900, 1900, 1900, 1900, 1900, 1900, 190...
## $ population <int> 1828697, 63592, 122931, 1311564, 1485053, 539700, 9...
```

```
#populatsioon kogu perioodi kohta
us_state_populations %>%
  select(state, population) %>%
  glimpse()
```

```
## Observations: 981
## Variables: 2
## $ state      <fct> Connecticut, Delaware, Georgia, Maryland, Massachus...
## $ population <int> 237655, 59096, 82548, 319728, 475199, 141899, 18413...
```

Võime muutujaid valida ka veeru nimes sisalduvate sõnamustrite abil

```
us_state_populations %>%
  #otsib veerge mis sisaldavad mustrit "pop" ja lõppevad tähtedega "ion"
  select(contains("pop"), ends_with("ion")) %>%
  glimpse
```

```
## Observations: 981
## Variables: 1
## $ population <int> 237655, 59096, 82548, 319728, 475199, 141899, 18413...
```

Veergude ümbernimetamine:

```
us_state_populations %>%
  rename(population=population, aasta = year, osariik=state) %>%
  glimpse
```

```
## Observations: 981
## Variables: 3
## $ aasta      <int> 1790, 1790, 1790, 1790, 1790, 1790, 1790, 1790, 179...
## $ osariik    <fct> Connecticut, Delaware, Georgia, Maryland, Massachus...
## $ population <int> 237655, 59096, 82548, 319728, 475199, 141899, 18413...
```

Muutujate loomine/ muutmine

Kui meil on vaja luua uusi muutujaid, siis me oleme seda juba teinud, kui arvutasime keskmise. Vahel on vaja muutujaid aga ümber arvutada. Näiteks tahame populatsiooni kujutada tuhandetes.

```
us_state_populations<-us_state_populations %>%
  #mutate loob uue muutuja
```

```
  mutate(population_thousand=population/1000)

head(us_state_populations)
```

```
##   year         state population population_thousand
## 1 1790   Connecticut     237655             237.655
## 2 1790      Delaware      59096              59.096
## 3 1790       Georgia      82548              82.548
## 4 1790      Maryland     319728             319.728
## 5 1790 Massachusetts     475199             475.199
## 6 1790 New Hampshire     141899             141.899
```

Tahame osariigi nimed sättida nii, et nad oleks väiketähtedega ja kokku kirjutatud. **Miks see kasulik on?**

```
#asendame tühikud alakriipsuga
us_state_populations$state=gsub(" ", "_", us_state_populations$state)
#teeme väiketähtdeks
us_state_populations$state=tolower(us_state_populations$state)
head(us_state_populations)
```

```
##   year         state population population_thousand
## 1 1790   connecticut     237655             237.655
## 2 1790      delaware      59096              59.096
## 3 1790       georgia      82548              82.548
## 4 1790      maryland     319728             319.728
## 5 1790 massachusetts     475199             475.199
## 6 1790 new_hampshire     141899             141.899
```

Nii, nüüd on vaja luua uus muutuja, mis on kategooriline muutuja (faktor), mis näitab perioodi 5-aastaste intervallidena.

```
us_state_populations$period<-cut(us_state_populations$year,
    seq(min(us_state_populations$year)-1,
        max(us_state_populations$year)+1,5))
#kontrollime, kas iga aasta kohta on üks intervall
table(us_state_populations$year, us_state_populations$period)[1:5,1:6]
```

```
##
##        (1789,1794] (1794,1799] (1799,1804] (1804,1809] (1809,1814]
##   1790          15           0           0           0           0
##   1800           0           0          20           0           0
##   1810           0           0           0           0          24
##   1820           0           0           0           0           0
##   1830           0           0           0           0           0
##
##        (1814,1819]
##   1790           0
##   1800           0
##   1810           0
##   1820           0
##   1830           0
```

Puuduvad väärtused

Puuduvad andmed on problemaatilised. Me saame kontrollida, kui paljudes veergudes on puuduvad väärtused (NA - not available).

```
colSums(is.na(us_state_populations))
```

```
##                year               state          population
##                   0                   0                   0
## population_thousand              period
##                   0                  52
```

Mida puuduvate väärtustega teha? Ideaalis tuleks need väärtused leida. Kui see pole võimalik, tuleks aru saada, miks need väärtused on puudu. Võib-olla saab need asendada keskmise mediaani, moodiga? Üks võimalus oleks luua kategooriline muutuja, mis näitab, kas väärtus oli puudu või mitte. Halvim valik oleks andmete eemaldamine!

```
#mis aastatel puudub period väärtus?
us_state_populations[is.na(us_state_populations$period),]$year
```

```
##  [1] 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010
## [15] 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010
## [29] 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010
## [43] 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010
```

```
#meie puhul on siin tegemist näpukaga.
```

```
#kui me tahame eemaldada read, kus on puuduvad väärtused, saab seda teha järgnevalt
us_state_populations[complete.cases(us_state_populations), ]
```

Andmete ühendamine

Meil on veel andmeid osariikide kohta:

```
df_states=read_csv('data/states.csv')
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   state = col_character(),
##   region = col_character(),
##   density = col_double(),
##   metro = col_double(),
##   waste = col_double(),
##   toxic = col_double(),
##   green = col_double(),
##   high = col_double(),
##   college = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
head(df_states)
```

```
## # A tibble: 6 x 21
##   state region    pop   area density metro waste energy miles toxic green
##   <chr> <chr>   <int>  <int>   <dbl> <dbl> <dbl>  <int> <int> <dbl> <dbl>
## 1 Alab~ South  4.04e6  52423  77.1    67.4 1.11     393 10500 27.9   29.2
```

```
## 2 Alas~ West   5.50e5 570374   0.960  41.1 0.910    991  7200 37.4   NA
## 3 Ariz~ West   3.66e6 113642  32.2    79   0.790    258  9700 19.6   18.4
## 4 Arka~ South  2.35e6  52075  45.2    40.1 0.850    330  8900 24.6   26.0
## 5 Cali~ West   2.98e7 155973 191.     95.7 1.51     246  8700  3.26  15.6
## 6 Colo~ West   3.29e6 103730  31.8    81.5 0.730    273  8300  2.25  21.9
## # ... with 10 more variables: house <int>, senate <int>, csat <int>,
## #   vsat <int>, msat <int>, percent <int>, expense <int>, income <int>,
## #   high <dbl>, college <dbl>
```

Ühendamiseks kasutame osariigi nime. Enne peame veenudma, et need klapiks kahes andmehulgas.

```
unique(df_states$state)
```

```
##  [1] "Alabama"              "Alaska"               "Arizona"
##  [4] "Arkansas"             "California"           "Colorado"
##  [7] "Connecticut"          "Delaware"             "District of Columbia"
## [10] "Florida"              "Georgia"              "Hawaii"
## [13] "Idaho"                "Illinois"             "Indiana"
## [16] "Iowa"                 "Kansas"               "Kentucky"
## [19] "Louisiana"            "Maine"                "Maryland"
## [22] "Massachusetts"        "Michigan"             "Minnesota"
## [25] "Mississippi"          "Missouri"             "Montana"
## [28] "Nebraska"             "Nevada"               "New Hampshire"
## [31] "New Jersey"           "New Mexico"           "New York"
## [34] "North Carolina"       "North Dakota"         "Ohio"
## [37] "Oklahoma"             "Oregon"               "Pennsylvania"
## [40] "Rhode Island"         "South Carolina"       "South Dakota"
## [43] "Tennessee"            "Texas"                "Utah"
## [46] "Vermont"              "Virginia"             "Washington"
## [49] "West Virginia"        "Wisconsin"            "Wyoming"
```

```
unique(us_state_populations$state)
```

```
##  [1] "connecticut"             "delaware"
##  [3] "georgia"                 "maryland"
##  [5] "massachusetts"           "new_hampshire"
##  [7] "new_jersey"              "new_york"
##  [9] "north_carolina"          "pennsylvania"
## [11] "rhode_island"            "south_carolina"
## [13] "southwest_territory"     "vermont"
## [15] "virginia"                "district_of_columbia"
## [17] "indiana_territory"       "kentucky"
## [19] "mississippi_territory"   "northwest_territory"
## [21] "tennessee"               "illinois_territory"
## [23] "louisiana_territory"     "michigan_territory"
## [25] "ohio"                    "orleans_territory"
## [27] "alabama"                 "arkansas_territory"
## [29] "illinois"                "indiana"
## [31] "louisiana"               "maine"
## [33] "mississippi"             "missouri_territory"
## [35] "florida_territory"       "missouri"
## [37] "arkansas"                "iowa_territory"
## [39] "michigan"                "wisconsin_territory"
## [41] "california"              "florida"
## [43] "iowa"                    "minnesota_territory"
## [45] "new_mexico_territory"    "oregon_territory"
```

```
## [47] "texas"                   "utah_territory"
## [49] "wisconsin"               "colorado_territory"
## [51] "dakota_territory"        "kansas_territory"
## [53] "minnesota"               "nebraska_territory"
## [55] "nevada_territory"        "oregon"
## [57] "washington_territory"    "arizona_territory"
## [59] "idaho_territory"         "kansas"
## [61] "montana_territory"       "nebraska"
## [63] "nevada"                  "west_virginia"
## [65] "wyoming_territory"       "alaska_territory"
## [67] "colorado"                "idaho"
## [69] "montana"                 "north_dakota"
## [71] "oklahoma_territory"      "south_dakota"
## [73] "washington"              "wyoming"
## [75] "hawaii_territory"        "utah"
## [77] "persons_in_the_military" "oklahoma"
## [79] "arizona"                 "new_mexico"
## [81] "alaska"                  "hawaii"
## [83] "puerto_rico"
```

Ei klapi, teeme väiketähtedeks ja asendame tühikud

```
df_states$state=gsub(" ", "_", df_states$state)
df_states$state=tolower(df_states$state)
unique(df_states$state)
```

```
##  [1] "alabama"              "alaska"               "arizona"
##  [4] "arkansas"             "california"           "colorado"
##  [7] "connecticut"          "delaware"             "district_of_columbia"
## [10] "florida"              "georgia"              "hawaii"
## [13] "idaho"                "illinois"             "indiana"
## [16] "iowa"                 "kansas"               "kentucky"
## [19] "louisiana"            "maine"                "maryland"
## [22] "massachusetts"        "michigan"             "minnesota"
## [25] "mississippi"          "missouri"             "montana"
## [28] "nebraska"             "nevada"               "new_hampshire"
## [31] "new_jersey"           "new_mexico"           "new_york"
## [34] "north_carolina"       "north_dakota"         "ohio"
## [37] "oklahoma"             "oregon"               "pennsylvania"
## [40] "rhode_island"         "south_carolina"       "south_dakota"
## [43] "tennessee"            "texas"                "utah"
## [46] "vermont"              "virginia"             "washington"
## [49] "west_virginia"        "wisconsin"            "wyoming"
```

Ühendame kakas andmehulka

```
df=merge(us_state_populations, df_states,by.x='state', by.y='state')
head(df)
```

```
##     state year population population_thousand      period region     pop
## 1 alabama 2010    4779736            4779.736        <NA>  South 4041000
## 2 alabama 1860     964201             964.201 (1859,1864]  South 4041000
## 3 alabama 1960    3266740            3266.740 (1959,1964]  South 4041000
## 4 alabama 1890    1513017            1513.017 (1889,1894]  South 4041000
## 5 alabama 1850     771623             771.623 (1849,1854]  South 4041000
## 6 alabama 1820     127901             127.901 (1819,1824]  South 4041000
```

```
##    area density metro waste energy miles toxic green house senate csat
## 1 52423   77.08  67.4  1.11    393 10500 27.86 29.25    30     10  991
## 2 52423   77.08  67.4  1.11    393 10500 27.86 29.25    30     10  991
## 3 52423   77.08  67.4  1.11    393 10500 27.86 29.25    30     10  991
## 4 52423   77.08  67.4  1.11    393 10500 27.86 29.25    30     10  991
## 5 52423   77.08  67.4  1.11    393 10500 27.86 29.25    30     10  991
## 6 52423   77.08  67.4  1.11    393 10500 27.86 29.25    30     10  991
##   vsat msat percent expense income high college
## 1  476  515       8    3627  27498 66.9    15.7
## 2  476  515       8    3627  27498 66.9    15.7
## 3  476  515       8    3627  27498 66.9    15.7
## 4  476  515       8    3627  27498 66.9    15.7
## 5  476  515       8    3627  27498 66.9    15.7
## 6  476  515       8    3627  27498 66.9    15.7
```

Kõikide osariikide kohta meil infot pole. Uurime, kas kõik osariigid on kogu algses andmehulgas olemas.

```
us_state_populations %>%
  group_by(year) %>%
  summarise(unique_states=length(unique(state)))
```

```
## # A tibble: 23 x 2
##     year unique_states
##    <int>         <int>
##  1  1790            15
##  2  1800            20
##  3  1810            24
##  4  1820            27
##  5  1830            28
##  6  1840            30
##  7  1850            36
##  8  1860            42
##  9  1870            47
## 10  1880            48
## # ... with 13 more rows
```

Apply jne